DOCUMENT RESUME

ED 128 417 TM 005 601

AUTHOR Steinheiser, Frederick, Jr.
TITLE A Bayesian Method for Maximizing Correct Mastery Classifications.
PUB DATE [Sep 75]
NOTE 24p.; Paper presented at the Annual Conference of the Military Testing Association (17th, Fort Benjamin Harrison, Indiana, September 15-19, 1975); Also included in TM 005 585

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Ability; Achievement; *Bayesian Statistics; *Classification; *Criterion Referenced Tests; Cutting Scores; *Mathematical Models; Military Personnel; Personnel Evaluation; *Probability; Psychometrics; Scores; *Testing; Test Interpretation
IDENTIFIERS Army; Mastery Tests

ABSTRACT

Summarizing work which is part of an Army research program on Methodological Issues in the Construction of Criterion Referenced Tests, the focus of this paper is on a Bayesian model, which gives the probability of correctly classifying an examiner as a master or as a nonmaster while taking into consideration the test length and the mastery cut-off score. Bayes' Theorem is a mathematical expression which allows the combination of information about the quality of the examinee population so as to produce a probabilistic estimate of mastery for a specific examinee. This approach can give the most accurate ability estimate for each examinee by using the fewest number of test items, provided that accurate estimates of the "quality parameters" have been made. A method of estimating these parameters from commonly available information is also explained. (Author/BW)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED128417

TM005 601

A Bayesian Method for Maximizing Correct Mastery Classifications

Frederick Steinheiser, Jr., PhD
Army Research Institute for the Behavioral and Social Sciences
Arlington, Virginia 22209

Introduction

This paper summarizes work which is part of an on-going research program at the Army Research Institute. The program, called Mettest--Methodological Issues in the construction of Criterion-Referenced Tests--is exploring and developing psychometric models for defining test standards and test lengths. The focus of this paper is upon a "Bayesian" model, which gives the probability of correctly classifying an examinee as a master or as a nonmaster while taking into consideration the test length and the mastery cut-off score.

Personnel assessment is typically made through the development and administration of tests, and the evaluation of test scores. The final desired output of a test for a given examinee is information which allows us to pinpoint his ability to do whatever is required by an objective. That is, we observe a test score and must then infer the ability of the examinee.

In Norm-Referenced Testing, an examinee's score is evaluated with respect to his position among all of the other scores in the examinee population. But in Criterion-Referenced Testing his score is evaluated with respect to his passing or not passing a particular instructional objective, independent of the scores obtained by others in the examinee population. A passing score indicates that he is a "master" of that particular instructional objective, and a failing score indicates that he is a "nonmaster."

Ideally, if an examinee's score on a test is above the minimal passing standard, he would be correctly classified as having "mastery ability." Ability assessment would therefore be based upon 100% correct classifications. But since we live in a less than ideal world, there will be variability due to the imperfections in the test construction, psychological variability (forgetting, guessing, individual differences) in the examinee population, and other unknown sources of error. Hence, sometimes a person who is really not a master (in the ideal world) will be classified as a master on the basis of his test score; and sometimes a person who really is a master will be unfortunate enough to be classified as a nonmaster. The following chart illustrates these four classification outcomes.

The views expressed in this paper are those of the author and do not necessarily imply endorsement by the U.S. Army.

		True Ability State: Master	Nonmaster
Classification Based Upon Test Score	Master	True positive	False positive
	Nonmaster	False negative	True negative

In order to approach the ideal classification accuracy, the probability of a True positive should be much greater than that for a False positive, and the probability for a True negative should be much greater than that for a False negative. The classification problem has now been cast into a decision-making framework, for which "Bayes' Theorem" may be used: we wish to obtain the probability of an examinee being in the Mastery Ability state given (conditional upon) his test score. Symbolically, this is expressed as $p(M1|T)$, where M1 refers to the mastery state, and T is the test score of that specific individual.

An Example Using Bayes' Theorem

Bayes' Theorem is a mathematical expression which allows us to combine information about the quality of the examinee population so as to produce a probabilistic estimate of mastery for a specific examinee. This approach will give the most accurate ability estimate for each examinee by using the fewest number of test items, provided that accurate estimates of the "quality parameters" have been made. (A subsequent section of this paper will show exactly how to estimate these quality parameters from data similar to that which you might have.) First, let's take a look at the mathematical expression, the estimates that we have to feed into it, and the output that it gives us:

$$p(M1|T) = \frac{p(T|M1)p(M1)}{[p(T|M1)p(M1) + p(T|M2)p(M2)]}$$

Here we assume that the 2 states of nature (master and nonmaster) are mutually exclusive and collectively exhaustive, and that T is the test score which is observed. We also assume that the test is dichotomously scored. A correct response is denoted "1", an incorrect response is denoted "0" and the total test score is simply the number of correct responses. What we seek to find is the term on the left, the probability that a given student is a master, having been given his test score. In order to find it, we need to have an estimate of the prior probability of mastery ($p(M1)$) in the population of students from which this student was drawn. The prior probability of mastery can be thought of as the proportion of students in the examinee population we think are masters. For example, if our instruction were very good the prior probability of

mastery would be high, and most of the students who completed the instruction should have mastered the objective. The actual number specified for the prior probability of mastery may be an informed guess based on experience or it may be based on the empirical results of tests given to previous classes of similar students.

We must also estimate the conditional probability of a certain test score given that the student who got that socre was a master. For example, if only one item were administered, the conditional probability of a score of one correct given that the student was a master is simply the probability that a master responds correctly. We may estimate this conditional probability empirically based on previous student groups, or we may provide a best guess as to how well masters perform, or this conditional probability may reflect a minimal standard of achievement. We shall show how the $p(M|T)$ will vary as a function of the prior expectations of the tester, number of test items, and conditional probabilities, $p(T|M)$, after an example to illustrate the computations.

Suppose that a student chosen at random from a trainee population was given a criterion-referenced test, and that he passed the test. Given the results of the test, what is the probability that the student is indeed a master of that particular course of instruction? In order to calculate the probability, we obtain the following information from the educational expert who administered the CRT: The probability that a master would obtain a passing score = .90, $(p(T|M1) = .90)$; the probability that a nonmaster would obtain a passing score = .05, $(p(T|M2) = .05)$; and the prior probability of randomly selecting a master from this trainee population is equal to .70, that is, we believe that 70% of this and similar previous trainee populations may be assumed to be composed of masters. Substituting these values into the formula;

$$p(M1|T) = \frac{.9 \times .7}{.9 \times .7 + .05 \times .3}$$

which equals .977. Hence, before the test score was available, the probability that this student was a master was .70, but after a passing score was observed, the probability that this person is a master has increased to .977. (The probability of this student's being a nonmaster, given the same passing score, $p(M2|T)$, would be equal to 1-.977 or .023).

Methods for Estimating "Quality Parameters" (Prior and Conditional Probabilities)

This model assumes that background information about students who previously took the test is available. This background information should lead to the accurate estimation of parameters that describe the quality of the examinee population. We need information to estimate the

prior probability of a randomly selected student being in one of the assumed mastery ability groups. We also need to be able to estimate the conditional probability that an examinee from a particular ability group would get an item right. For purposes of illustration, let's assume that 100 examinees produced the following distribution of scores on a five item test:

Number of items correct	Frequency	p(Correct)
5	30	1.0
4	30	.8
3	00	.6
2	10	.4
1	20	.2
0	10	0.0

For this particular set of data, it seems reasonable to postulate two ability groups. Note that 60 examinees got either 4 or 5 test items correct, and that a total of 40 examinees got 0, 1, or 2 items correct. No one got 3 items correct. Hence, this bimodal distribution of scores strongly suggests that we may set the prior probability of mastery equal to .6, and the prior probability of nonmastery equal to .4. Symbolically, $p(M1) = .6$, and $p(M2) = .4$.

We also need to estimate the conditional probability that a correct response is made to an item, given the particular mastery or ability state. Symbolically, we seek $p(x=1|M)$. There are several ways to compute this parameter value. Unfortunately, each method produces a unique value.

We could take the average proportion correct for the mastery group. For the present data, this would produce $\frac{30 \times 1.0 + 30 + .8}{60} = .9$.

We could also take the lowest score in the mastery group, which in this case is .8.

We could also take the desired (or "standard") score required for the demonstration of mastery, which need not necessarily be observed. This value for the present data could be set at .70, or .7x, or .82, etc.

This variety of estimated values should not be distressing, since it allows the examiner to introduce his own requirements into the selection. The important thing is that his choice should be close to at least one of the logically derived estimates.

The Bayesian Mathematical Model

In order to generalize the Bayesian approach to a wide variety of applications in personnel assessment, two additions must be made to the previously described formula. These additions are the number of trials or items on the test (N), and the number of hypothesized mastery ability states (S). The derivation of the general formula to meet these goals was originally presented by Hershman (1971):

$$p(Mi|T) = \frac{\prod_{j=1}^{N} p(Mi|t_j)}{p(Mi)\sum_{i=1}^{S-1}\frac{\prod_{j=1}^{N} p(Mi|t_j)}{p(Mi)^{S-1}}}$$

In this formula, $p(Mi|tj)$ equals the conditional probability of a person in the ith mastery state getting the jth test item correct; $p(Mi)$ is the prior probability of the representation of the ith mastery state in the student population (the % of students who are estimated to be in the ith mastery state); and $p(Mi|T)$ is the conditional probability of a particular student being classified as being in the ith mastery state given his total test score. A computational example showing how the formula is applied for three mastery states is given in the Appendix.

This formula was not used to analyze any "real" test score data. Rather, selected values of the various parameters were systematically manipulated in order to determine their influence on the probability of mastery classification. Hence, the results are from a computer simulation of idealized data and should serve to emphasize the relative effects of each of the parameters.

The two parameters that estimate the quality of the examinee population, the prior probability of selecting a master from the population, and the conditional probability of a known master and of a known non-master getting a randomly selected item correct, have already been discussed. Basically, $p(M1)$ reflects the proportion of masters in a group of examinees (and the level of training), whereas $p(1|M1)$ and $p(1|M2)$ will be high if the test is easy and lower if the test is difficult. The ideal criterion referenced test should provide a high probability for the former, and low probability for the latter.

Two other parameters, the minimal passing standard (the per cent of all the items that were answered correctly) and the test length are interrelated. By way of analogy, consider the minimal passing standard for deciding that a coin is biased to be 70%. That is, if heads (or

tails, for that matter) come up on 70% of the tosses, we would evaluate the coin as being unfair. Note that the 70% figure is arbitrary. We could have set the standard at 65%, or 75%, or 80%, etc. Now how many tosses (test items) do we want to observe? 10? 50? 100? 1,000? If we observe 7 heads out of ten tosses, and 700 heads out of 1,000 tosses, does the probability of the fairness of the coin remain the same? It should be intuitively obvious (and it can be easily demonstrated by means of the binomial distribution) that the minimal passing standard interacts with test length. The probability that a coin is fair when 7 heads out of ten tosses are observed is much greater than when 700 heads out of 1,000 tosses are observed--even though the "70%" criterion was strictly maintained! Values of 60%, 70%, and 80% correct were used in the current simulation. The number of trials or items (N) took on values of 5, 10, 20, and 40.

The final parameter which was manipulated in the model is the number of assumed mastery states. It may be overly simplistic to assume that the world is divided into just two dichotomous and mutually exclusive states, of mastery and nonmastery. Perhaps there are varying degrees of mastery, ranging from "complete" to "partial" to "totally incomplete." The present model is able to handle any number of assumed mastery states.

The model makes the following two important assumptions concerning the nature of the test from which the data are derived: (1) The test measures a unidimensional latent trait or unitary skill; (2) Test items or trials are equally difficult for a given ability. An elaboration of the basic model can easily be made to include test items with varying degrees of difficulty.

Changes in p(M|T) Assuming Two Mastery States

The fundamental purpose of the present study was to investigate how the probability of mastery classification changes as a function of the simultaneous manipulation of up to four parameters (independent variables). The scope of the study is not exhaustive, since only several were used. However, some general trends do seem to emerge as can be seen in the following figures.

Figures 1, 2, and 3 show the results of applying the model to a situation in which only two mastery groups (mastery and nonmastery) have been hypothesized. The data points represent the probability that a trainee is a master, given (conditional upon) his total test score, P(M|T). The curvature of each line shows how the P(M|T) changes as a function of variations in the prior expectation of mastery, the % correct items observed, the conditional probabilities of both a master and a nonmaster responding correctly to an item, and the number of items comprising the test.

Figure 1 represents a testing situation in which the training was of extremely high quality, since the proportion of masters in the trainee population was assumed to equal 0.9. That is, p(M1) = 0.9. Figure 1A portrays the situation in which both masters and nonmasters have attained a rather high degree of proficiency, since the probability of a master responding correctly to any given item is 0.9, and the probability of a nonmaster responding correctly is 0.6. If a person scored 80% on a five item test, the probability that he is a master is approximately .91. This probability drops to .65 if a 60% score on five items (3 out of 5 correct) were obtained. Note that when the test length is increased to 40 items, an 80% score (32 correct) produces a .99 probability of mastery. However, a score of 60% (24 correct) yields an essentially zero probability of mastery. The effect of the test length variable on classification accuracy is dramatic: if the p(M|T) had to be at least 0.5 for a person to be called a master, then scores of 60% on a five item test would lead to mastery classification. But a 60% score on a 40 item test would lead to nonmastery classification.

Figure 1A also illustrates the effect of "prior beliefs" on p(M|T). Intuitively, one might suppose that the chances were much higher that a person who obtained a score of 60% (even from a 5 item test) came from a population whose probability of correctly answering an item was 0.6 than from a population whose probability of answering an item correctly was 0.9. However, the relative proportions of the two groups (expressed as prior belief in mastery and nonmastery, or p(M1) = .9 and p(M2) = .1, respectively) are such that the probability of a person being in the mastery state is approximately 0.65 for a score of 3 correct (60%) on a 5 item test. Only by increasing the number of test items can the strong prior bias in favor of the mastery decision be reversed. Figures 2A and 3A show what happens when prior beliefs are not so heavily biased in favor of mastery. In neither case is the probability of being in the mastery state above 0.5 for scores of less than 80%. But Figure 1A suggests that when prior beliefs heavily favor one group over the other, longer length tests should be used. Otherwise, the amount of data may not be sufficient to force a change in the originally held prior beliefs.

The effect of changing the prior beliefs concerning the proportion of masters and nonmasters in the examinee population while holding all other parameters constant can be seen by comparing corresponding graphs A, B, C, and D in Figures 1, 2, and 3. As the prior beliefs approach equiprobability (where p(M1) = p(M2) = 0.5), more items are required to maintain a given level of confidence that a person is either a master or nonmaster. The inability to postulate strong prior beliefs must be compensated for by increasing the test length in order to maintain a constant classification accuracy.

The effect of changing the probability of a correct response, p(1|Mi), can be seen by comparing graphs A, B, C, and D for Figures 1, 2,

and 3. For example, the only difference between Figure 1A and Figure 1B is that the $p(1|M1)$ changes from 0.9 to 0.8, all other parameters being held constant. (This change might reflect a lower level of required proficiency, and hence less training, for Graph B than for A. Or perhaps previous test results indicate that masters of the instruction respond to items with a probability of correct response equal to 0.8 rather than 0.9.) In any case, the effect of this small change in the $p(1|M1)$ on the $p(M|T)$ is readily apparent. For any test length or observed test score, the probability of being in the mastery state is greater in Graph B than in A. This shift is most obvious for the 70% observed correct curve. Notice that $p(M|T)$ on Graph A for an observed score of 70% (28 out of 40 correct) is approximately 0.04. However, the value for $p(M|T)$ in Graph B for 70% of a 40 item test correct is 0.87.

The main reason for this abrupt change from Graph A to B (in Figures 1, 2, and 3) is the lowered requirement for mastery, from 0.9 to 0.8. The probability that "0.9 persons" score only 70% correct on long tests is relatively low. But when masters are defined as those trainees who come from a population with a probability of responding correctly equal to 0.8, the probability of their scoring 70% on a long test is high. One of the most difficult jobs for an instructional designer is to describe the level of capability required of graduates and the level of capability actually achieved. Comparison of these graphs indicates the magnitude of the effect that these specifications can have on the classification of trainees.

Graphs C and D of Figures 1, 2, and 3 further illustrate the effect of variations in the probability of correct responses. The only difference between Graphs B and C is that the probability of a correct response from a nonmaster decreases from 0.6 to 0.5. The effect of this decrease in correct response probability from a nonmaster is to lower the likelihood of a nonmaster achieving a test score of at least 70%, which also increases the probability that a person achieving a high % score is in the mastery state. Finally, Graph D portrays an extreme case in which neither masters nor nonmasters are responding at particularly high levels. However, the level of performance for nonmasters is so low (0.4), that even for observed scores of 60% the probability of being in the mastery state exceeds 0.8 for all test lengths, except for 5 and 10 items in Figure 2, and 5, 10, and 20 items in Figure 3.

Further detailed analysis of these figures is not included in this paper. In comparing the twelve graphs against each other, note the magnitude of the changes in $p(M|T)$ when small changes have been made in the prior beliefs, in the correct response probabilities, and in the percent correct observed responses. The implication is that extreme care must be taken when specifying parameters in a Bayesian approach to testing and decision making. If the parameters are realistic, great savings in

testing time and expense, and increased confidence in decision making are possible (Novick & Lewis, 1974). However, if the parameters are not realistic, there is a very real danger of misclassifying many examinees. The next section of this paper deals with an elaboration of the model to three mastery states, thus helping to quantify sources of classification error.

Elaboration to Three Mastery States

Figures 4, 5, 6, and 7 represent cases for which three mastery states have been hypothesized. In figures 4 and 6 the probability of a correct response for a person assumed to be in mastery state M1 equals 0.8, for mastery state M2 this probability is 0.6, and for mastery state M3 it is 0.5. These values could correspond to the situation in which the non-mastery group was divided in half. That is, those persons whose probability of getting any given item correct is 0.5 (comprising mastery state M3) would need extensive retraining; whereas those whose probability of getting any given item correct is 0.5 (comprising mastery state M3) would need extensive retraining; whereas those whose probability is 0.6 (comprising mastery state M2) would merely need selective retraining. People in mastery state M1 have a probability of 0.8 for making a correct response, and may therefore be considered as "masters" who have successfully passed training.

For Figures 5 and 7 the corresponding probabilities of a correct response for people in mastery states M1, M2 and M3 are 0.9, 0.8, and 0.6, respectively. These probabilities might describe a situation in which the mastery group was dichotomized, perhaps in an attempt to identify those students who had achieved an exceptionally high level of proficiency, i.e., $p(1|M1) = 0.9$.

In Figures 4 and 5 the prior probability (or assumed proportion) of examinees in each mastery state are: $p(M1) = 0.5$, $p(M2) = 0.3$, and $p(M3) = 0.2$. In Figures 6 and 7 the corresponding prior probabilities are 0.25, 0.50, and 0.25, respectively. The prior values in Figures 4 and 5 display a bias towards higher levels of mastery (50% of the examinees are assumed to be type M1 masters), whereas the bias in Figures 6 and 7 is towards the intermediate level of mastery (50% of the examinees are assumed to be type M2 masters).

A detailed analysis of Figures 4 and 5 will provide the basis for an interpretation of Figures 6 and 7, which is an exercise left to the reader. The three graphs labeled A, B, and C represent the probability that an individual is in mastery state M1, M2, and M3, respectively.

Graph A shows the probability that an individual is in mastery state M1 given observed scores of 60%, 70%, and 80% correct on 5, 10, 20, and 40 item tests. Thus, for an observed score of 4 out of 5 correct, the

probability that this person is in mastery state M1 is about 0.65. But if this same person got a score of 32 out of 40 (still 80% correct), the probability that he is an M1 master jumps to 0.98. These results are similar to those obtained when two mastery groups were hypothesized, and again illustrate the effect of increasing test length on the level of confidence in the mastery classification $p(M|T)$.

The probability of being in mastery state M2 given observed scores is plotted in Graph B. If a person got 4 out of 5 correct, the probability of being in state M2 is about 0.25. However, if he got 32 out of 40 correct (still 80% correct), this probability plummets to 0.02. Finally, using these same test score values, Graph C shows that the probability of being a type M3 master is 0.10 for 4 out of 5 correct, and nearly zero for 32 out of 40 correct. This result makes intuitive sense, because there is only 20% of type M3 (non)masters in the examinee population, and the probability of their getting any item correct is only 0.50, which is a long way from 80% observed correct.

Notice that for any given test length and percent correct, the sum of the probabilities of being in states M1, M2, and M3 equals 1.0. Comparison of Graphs A, B, and C shows that when either 70% or 80% of the items for any test length are correctly answered, the probability of being in state M1 is greater than the probability of being in either state M2 or M3. That is, both the 70% and 80% curves are higher in Graph A than in either Graph B or C. For an observed score of 60% the probability of being in state M2 is greater than for M1 or M3. The probability of being in state M3 is rather low for all values of test length and percent correct observed in this particular example.

In Figure 5 the interrelationship between test length and three hypothesized mastery states becomes even more apparent. For example, Graph A shows that the probability of being in state M1 for 80% correct on a 5 item test is about 0.48. The probability of being in state M2 (shown in Graph B) for 80% correct on a five item test is about 0.36. There is thus a greater chance that a person whose score is 4 out of 5 is in M1 ($p(M1|T) = 0.48$), instead of M2 ($p(M2|T) = 0.36$) or M3 ($p(M3|T) = 0.16$). However, if a score of 80% correct were observed on a 40 item test, the graphs indicate that a much different decision would be appropriate. In this case, $p(M1|T)$ equals 0.21, $p(M2|T) = .78$, and $p(M3|T) = 0.01$. Hence, people scoring 32 out of 40 correct should be classified as type M2 masters. Also note that a score of 60% for any test length implies that these people should be placed in the M3 state.

For the data used in Figure 5, the probability of finding M1 type masters is overall quite low. Instead, for the levels of achievement demonstrated by obtained scores of 60%, 70%, or 80%, it is more likely that such scores were produced by people in mastery states M2 ($p(1|M2) = 0.8$) and M3 ($p(I|M3) = 0.6$).

Test Length and Misclassification Error

One of the most important questions that must be answered in designing a training evaluation program is: What is the probability of falsely classifying a person on the basis of a given observed score? It is also possible to turn the question around and ask: How long must a test be, and what score is required for classification decisions to be made with some specified lower limit of misclassification?

Figures 8 and 9 demonstrate how the Bayesian model can be used to answer the above questions. Assuming that the prior and conditional probabilities are realistic and fixed, the important variables are then test length and cutting score. Suppose that $p(M1) = 0.9$, $p(M2) = 0.1$, $p(1|M1) = 0.9$, and $p(1|M2) = 0.6$ as in Figure 8. In this example, the prior belief that an untested trainee is a master is very high, $p(M1) = 0.9$. A reasonable question might therefore be: What score must be observed such that a nonmastery decision can be made with at least 90% confidence? In other words, what data are required to force a reversal in the prior belief?

To be 90% confident of a nonmastery decision, $p(M2|T)$ must be equal to at least 0.90. Since the sum of $p(M1|T)$ and $p(M2|T)$ equals 1.0, $p(M1|T)$ must therefore not be greater than 0.10. Referring to Figure 8, a horizontal line crossing the ordinate at 0.10 can be drawn. This line crosses the curve for a five item test at a point corresponding to 26% correct. The next lowest possible test score is one correct (20%), so the decision rule is that all persons scoring one correct or less should be considered nonmasters. The point on the ordinate corresponding to 20% correct on the five item test is about 0.05. Hence, the final decision rule states that nonmastery decisions based on an observed score of one correct out of five can be made with 95% confidence (1.00 - 0.05 = 0.95). For observed scores lower than the cutoff score the confidence in making a correct decision must increase. Continuing with the present example, the $(pM1|T)$ if zero correct are observed is virtually equal to zero. Hence, those persons who get no items right may be classified as M2 type nonmasters with nearly 100% confidence.

A similar analysis applied to the 40 item test curve indicates that the cutting score should be about 73% correct. The next lowest possible score to 73% is 70%, which yields exactly 28 correct out of 40 items. The probability of mastery given an observed score of 28 correct is about 0.04. At such a low value of $p(M1|T)$ the chances for misclassification using a five item test and a 40 item test are almost the same. However, the observed percent correct at which the nonmastery decision is made for the two tests is 20% on the five item test and 70% on the 40 item test. Superficially, two tests of different lengths would seem to produce the same decision outcome, and that longer tests may not really be necessary for reducing classification error.

In order to appreciate the benefits gained by using longer length tests, the entire curve must be examined. Note that at 80% correct the five item test yields $p(M1|T)$ to equal 0.92. This result suggests that, on the average, 8% of the mastery decisions will be in error. For the 40 item test, the probability of mastery given 80% correct is about 0.99. That is, there is only about 1% chance of misclassification error. A test that distinguishes sharply between masters and nonmasters is one in which the probability of mastery is close to either 0.0 or 1.00 for most obtained scores. On such tests there is only a small region in which classification error is large. For example, in Figure 8 for the 40 item test, the region where $p(M1|T)$ is greater than 0.1 and less than 0.9 extends from 71% to 77% correct. This means that the probability of misclassifying a person will exceed 0.10 only when observed scores range from 71% to 77% correct. In contrast, the region of the five item test curve for which $p(M1|T)$ is greater than 0.10 and less than 0.9 extends from about 26% to about 79%. Hence, there is a much larger region of the curve for which the probability of misclassification exceeds 0.10. Obviously, if classification accuracy is to be maximized over the entire range of possible test scores, then longer tests are required. Ideally, a very long test would produce a step function, for which all values of possible scores approach either 0.0 or 1.0.

Figure 9 can be analyzed in a manner similar to that for Figure 8. However, Figure 9 has one outstanding characteristic that merits special attention. If nonmastery decisions must be made with 90% confidence, and a horizontal line at $p(M|T) = 0.1$ is drawn, the line does not intersect the curve for the five item test. This means that it is not possible to classify a nonmaster with 90% confidence if a five item test is used, given the parameters used in Figure 9. If resource or time constraints are such that no more than five items may be given, and if the parameter values used in Figure 9 are realistic, and if 90% confidence for mastery decisions are required, then there is no reason to test. Testing is irrelevant because no matter what score is observed, including zero correct, the decision rule compels a mastery decision to be made. In fact, for the present values, the probability of mastery given zero correct, is equal to 0.21. This simply means that if persons obtaining a score of zero are classified as nonmasters, 21% of them will be misclassified, on the average.

Conclusions

The implications of the results from this simulation experiment stress the practical importance of test length, criterion scores, and accurate estimates of examinee quality in making optimal mastery classifications. If the assumptions of the model have been met, and if accurate parameter estimates have been made, then a Bayesian method is optimal in the sense of providing more accurate estimates of mastery classification with the least number of test items.

In the typical situation for personnel assessment, the examiner will have some degree of control over the values of the parameters. His estimates of the prior probability of mastery will depend upon the goodness of the information he can obtain about previous examinee populations' scores. His estimates of the conditional probabilities can be made by several equally justifiable and logical procedures. In any case, informed subjective judgment is absolutely essential.

The criterion for minimal mastery, expressed as some percent correct of the total number of test items, is explicitly under the examiner's control. In some testing situations, he may deem 70% correct as a minimal passing score, whereas when more critcal skills are involved, he may want to observe at least 80% correct before calling an examinee a master. But as the model demonstrates, test length interacts with per correct required for "mastery" decisions. Specifically, as test length increases, classification accuracy increases, even when the same percent correct is maintained. In performance-based tests for example, where the cost of each item could be very high (such as field artillery or tank gunnery), the examiner is obliged to use the minimum number of trials, and so the minimal percent correct mastery criterion should be increased accordingly. Finally, the model has demonstrated that testing may be irrelevant in making mastery classification decisions if test length does not exceed some minimal number of items.

Appendix: A Computational Example for Three Mastery States

The following example illustrates the computations necessary for processing data with the Bayesian model. The values chosen for this example correspond to Figure 4. Assume that there are three states of mastery, and unequal prior probabilities for these three states. The educational decision-maker must provide estimates for the prior probabilities of master, p(Mi). For this example let us assume the values to be: $p(M1) = .5$; $p(M2) = .3$; and $p(M3) = .2$. He must also provide estimates for the conditional probability of getting any given test item right, given each mastery state. The following values will be used as the conditional probability of getting an item right given a mastery state: $p(1/M1) = .8$; $p(1/M2) = .6$; $p(1/M3) = .5$. The conditional probabilities of getting an item wrong given a mastery state are: $p(0/M1) = .2$; $p(0/M2) = .4$; and $p(0/M3) = .5$.

First we need to calculate the probability that an item is answered correctly. For the overall population,

$$p(tj = \text{correct}) = \sum_{i=1}^{S} p(Mi)p(tj = \text{correct}/Mi) = (.5)(.8) +$$

$$(.3)(.6) + (.2)(.5) = .68.$$

Likewise,

$$p(tj = \text{wrong}) = \sum_{i=1}^{S} p(Mi)p(tj = \text{wrong}|Mi) = (.5)(.2) +$$

$$(.3)(.4) + (.2)(.5) = .32.$$

We also need to obtain the set of conditional probabilities for the different mastery states given that an individual item was responded to either correctly or wrongly. The general equation is:

$$p(Mi|tj) = \frac{p(Mi)p(tj|Mi)}{p(tj)}$$

Substituting the above values yields: $p(M1|tj = \text{correct}) = (.5)(.8) \div .68 = .588$; $p(M2|tj = \text{correct}) = (.3)(.6) \div .68 = .265$; and $p(M3|tj = \text{correct}) = (.2)(.5) \div .68 = .147$. (Note that the sum equals 1.0.) Finally, $p(M1|tj = \text{wrong}) = (.5)(.2) \div .32 = .3125$; $p(M2|tj = \text{wrong}) = (.3)(.4) \div .32 = .375$ and $p(M3|tj = \text{wrong}) = (.2)(.5) \div .32 = .3125$. If 6 items were answered correctly on a 10 item criterion-referenced test, the following $\prod_{j=1}^{N} p(Mi|tj)$ values result: $M1 = 3.9 \times 10^{-4}$; $M2 = 6.8 \times 10^{-6}$; $M3 = 9.6 \times 10^{-8}$.

Finally, the general Bayesian formula yields the conditional probability for each mastery state given the total test score. For example,

$$p(Mi|T) = \frac{(3.9 \times 10^{-4})}{(.5)^9 \left[\frac{(3.9 \times 10^{-4})}{(.5)^9} + \frac{(6.8 \times 10^{-6})}{(.3)^9} + \frac{(9.6 \times 10^{-8})}{(.2)^9}\right]} = .272.$$

Similar calculations yield $p(M2|T) = .473$ and $p(M3|T) = .254$.

References

Hershman, R.L. A rule for the integration of Bayesian opinions. Human Factors, 1971, 13, 255-259.

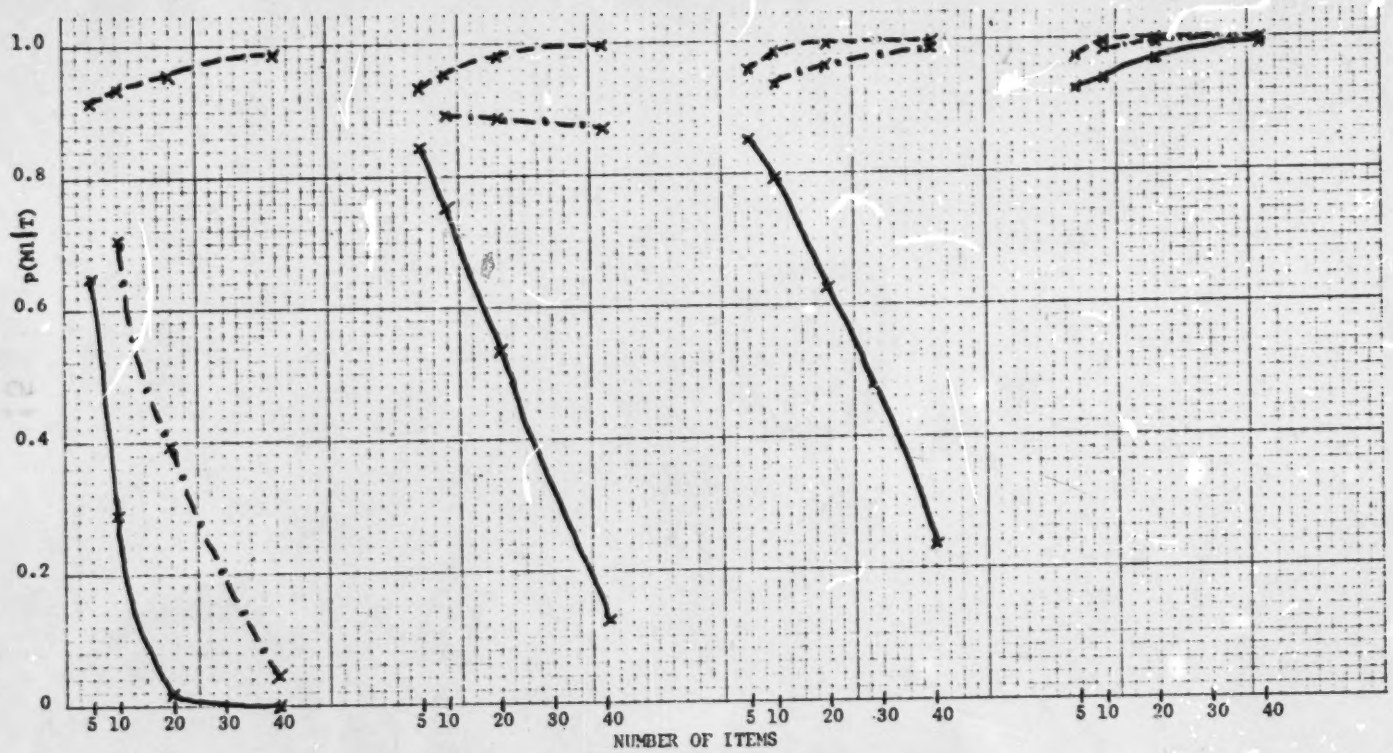
Novick, M.R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.), Center for the Study of Evaluation Monograph Series in Evaluation III: Problems in criterion-referenced measurement. Los Angeles: U.C.L.A. Center for the Study of Evaluation, 1974.

FIG. 1.

P(M1) = .9 P(M2) = .1

60% Correct ———— 70% Correct —·—·—·— 80% Correct — — — —·—

A	B	C	D
P(1|M1) = .9	P(1|M1) = .8	P(1|M1) = .8	P(1|M1) = .7
P(1|M2) = .6	P(1|M2) = .6	P(1|M2) = .5	P(1|M2) = .4



618

FIG. 2.

P(M1) = .7 P(M2) = .3

60% Correct ——————— 70% Correct — — — · — · — · — 80% Correct — — — — —

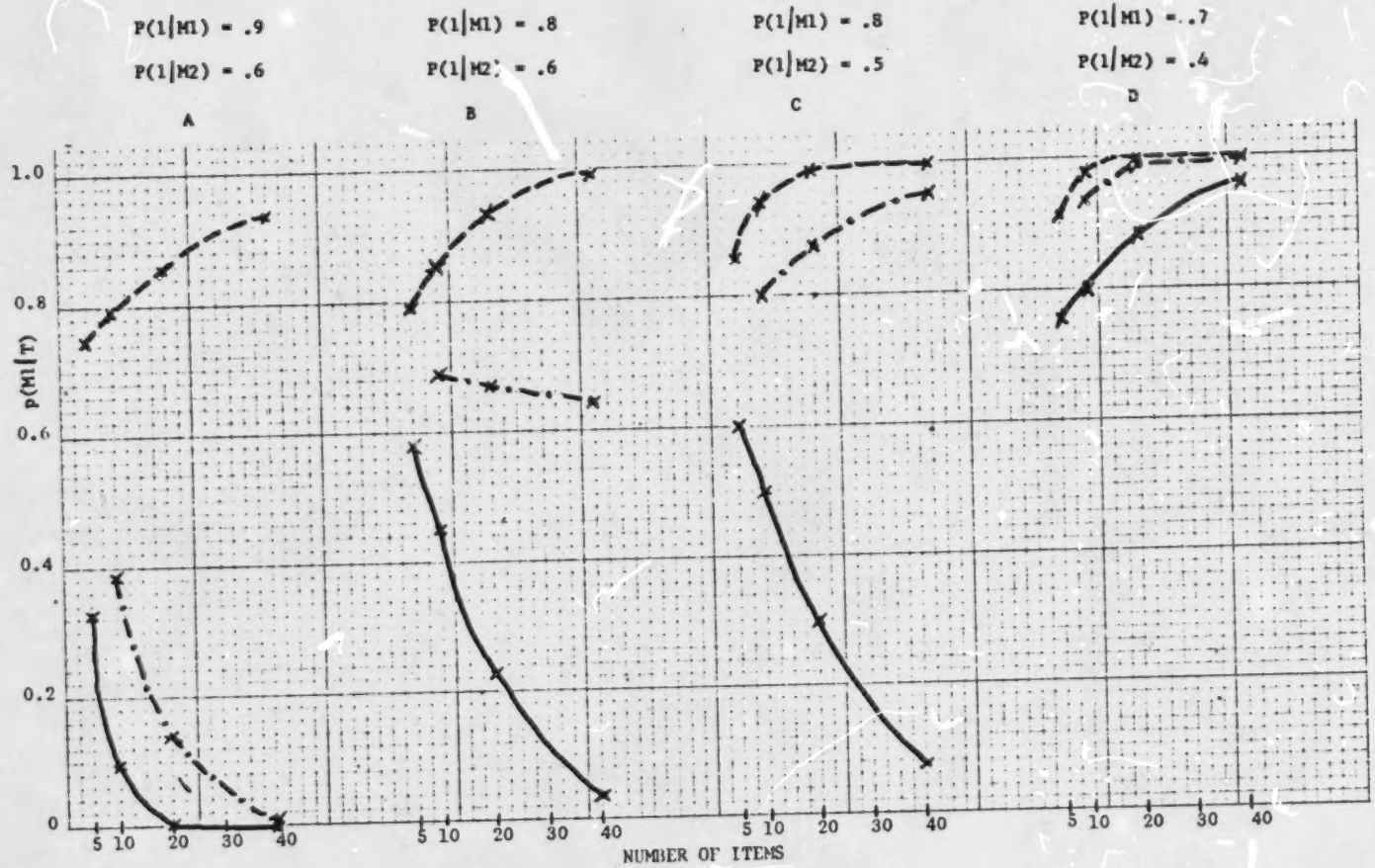


FIG. 3.

P(M1) = .5 P(M2) = .5

60% Correct —————— 70% Correct —·—·—·— 80% Correct — — — —

A	B	C	D
P(1|M1) = .9	P(1|M1) = .8	P(1|M1) = .8	P(1|M1) = .7
P(1|M2) = .6	P(1|M2) = .6	P(1|M2) = .5	P(1|M2) = .4

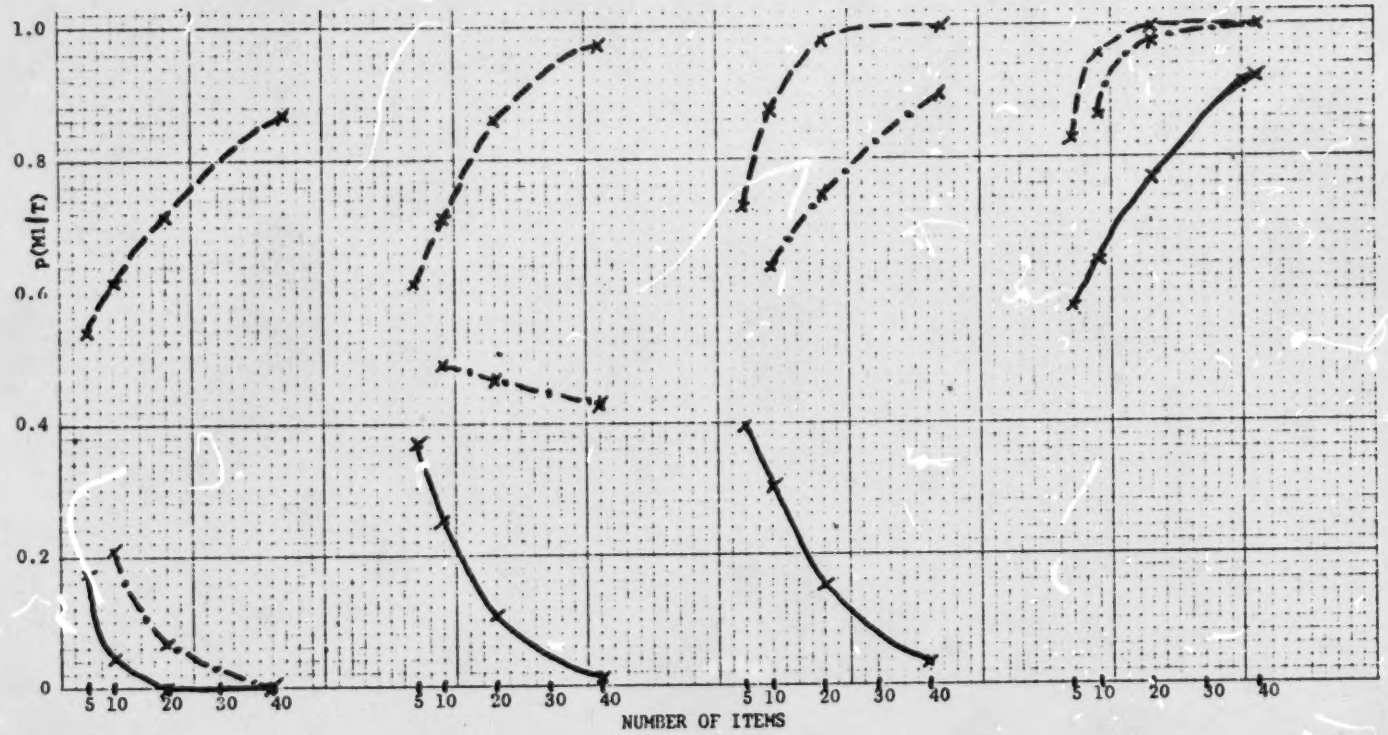


FIG. 4.

P(M1) = .50 . P(M2) = .30 P(M3) = .20

60% Correct ———— 70% Correct —·—·—·— 80% Correct — — — —

P(I|M1) = .8, P(I|M2) = .6, P(I|M3) = .5

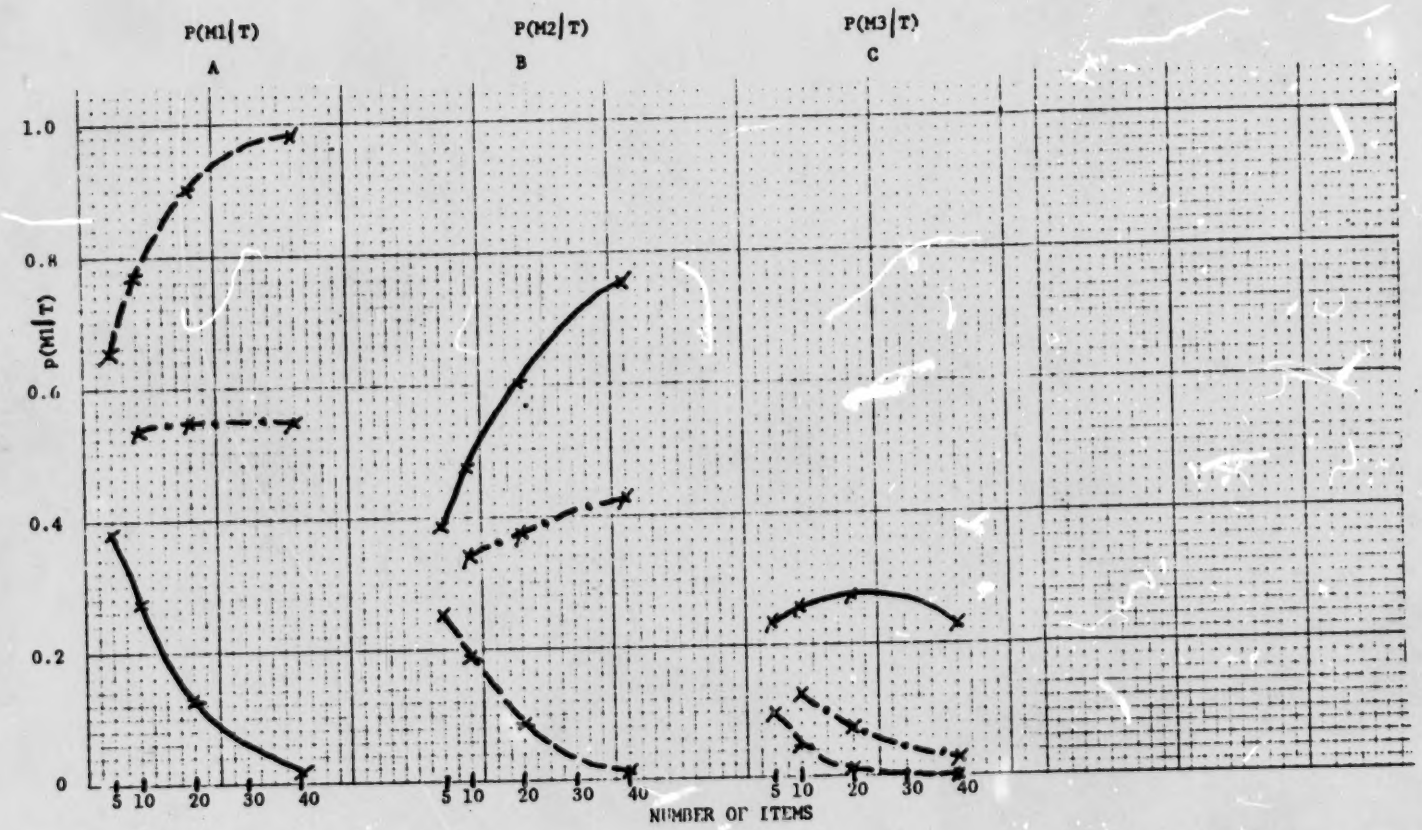
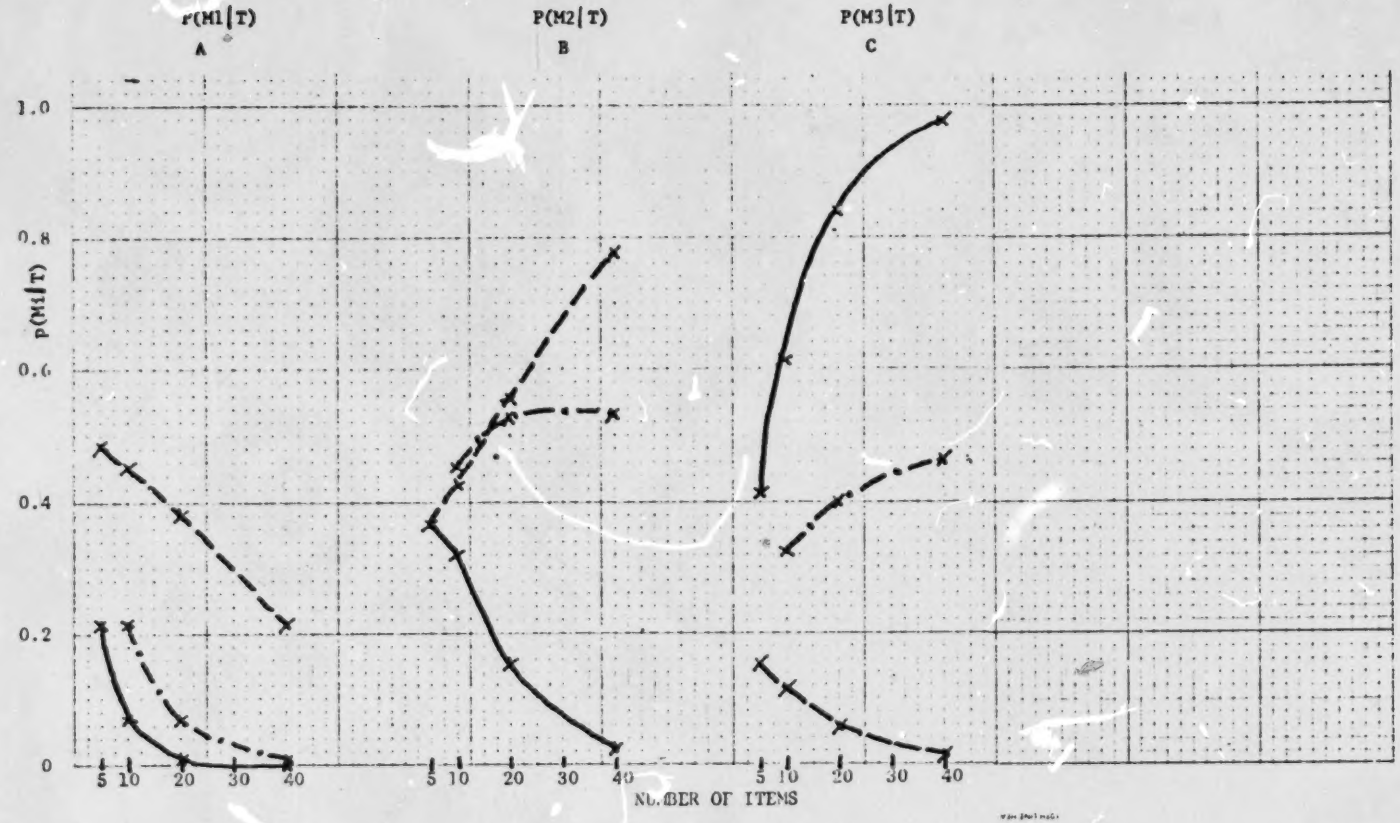


FIG. 5.

P(M1) = .50 P(M2) = .30 P(M3) = .20

60% Correct ——————— 70% Correct —·—·—·—·— 80% Correct — — — — —

P(1|M1) = .90, P(1|M2) = .80, P(1|M3) = .60



622

FIG. 6.

P(M1) = .25 P(M2) = .50 P(M3) = .25

60% Correct ——————— 70% Correct —·—·—·— 80% Correct — — — — —

P(1|M1) = .8, P(1|M2) = .6, P(1|M3) = .5

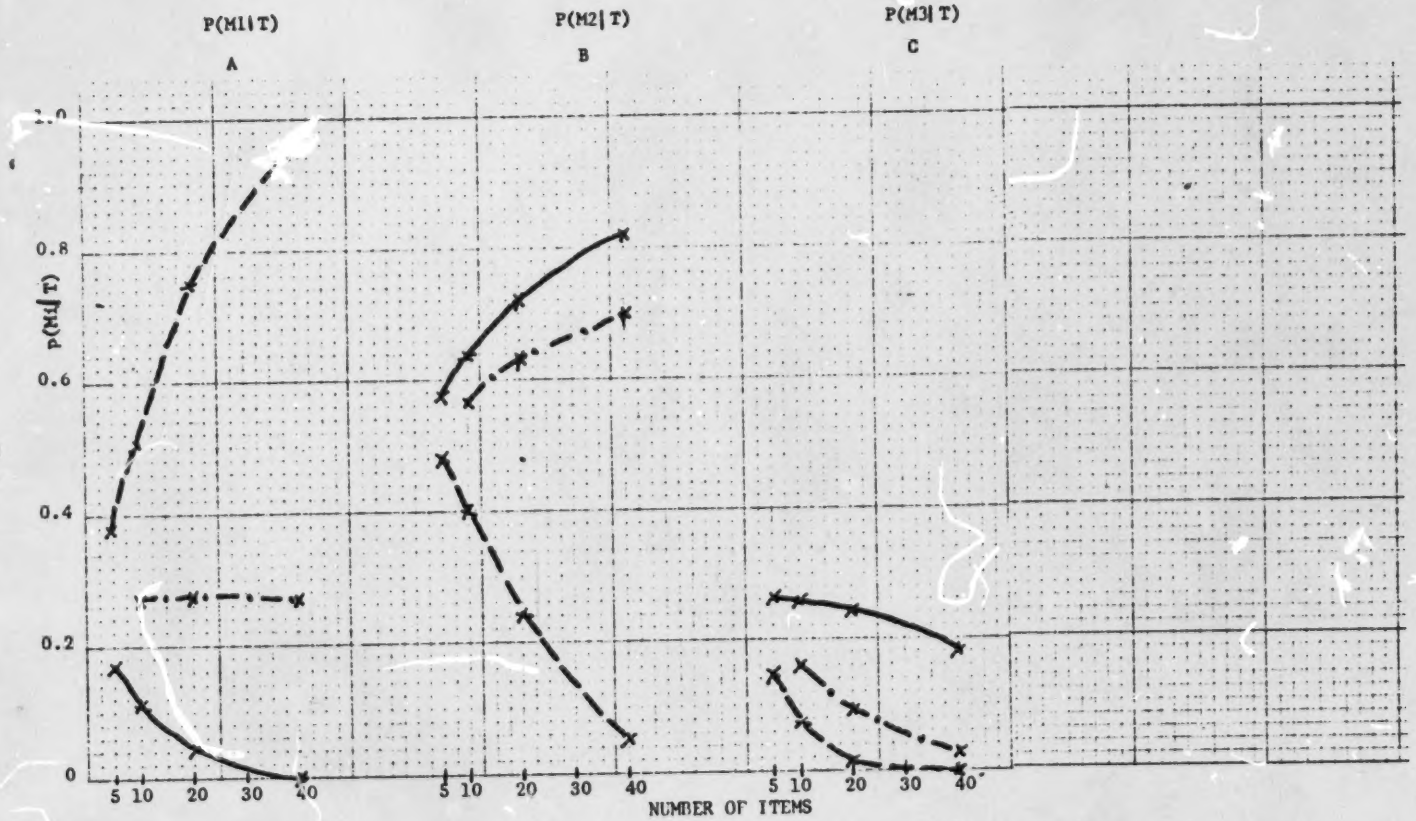
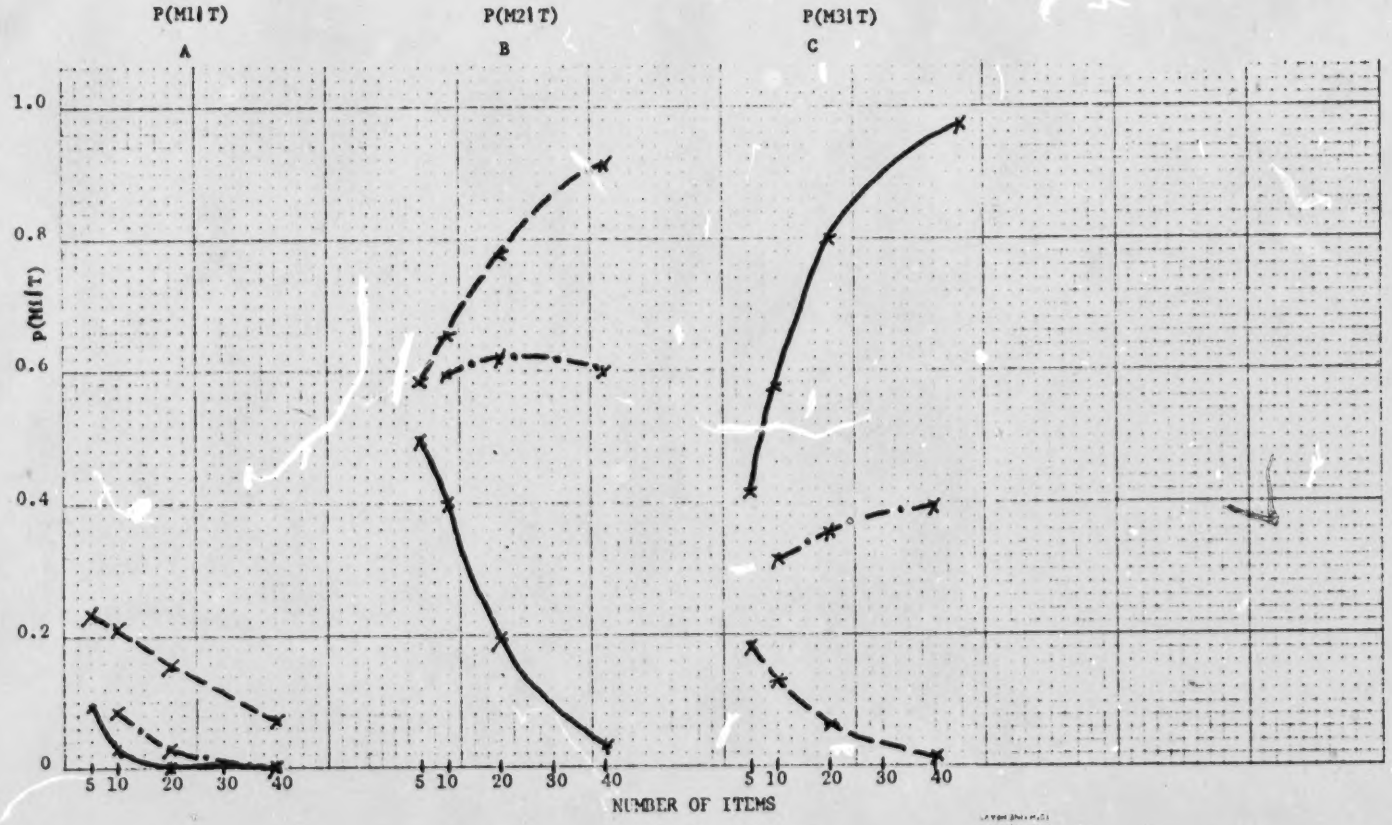


FIG. 7.

P(M1) = .25 P(M2) = .50 P(M3) = .25

60% Correct ——————— 70% Correct —·—·—·—·— 80% Correct — — — — — —

P(1|M1) = .90, P(1|M2) = .80, P(1|M3) = .60



624

FIG. 8.

P(M1) = .9 P(M2) = .1

P(1|M1) = .9, P(1|M2) = .6

5 Item Test - - - - - - 40 Item Test ——————

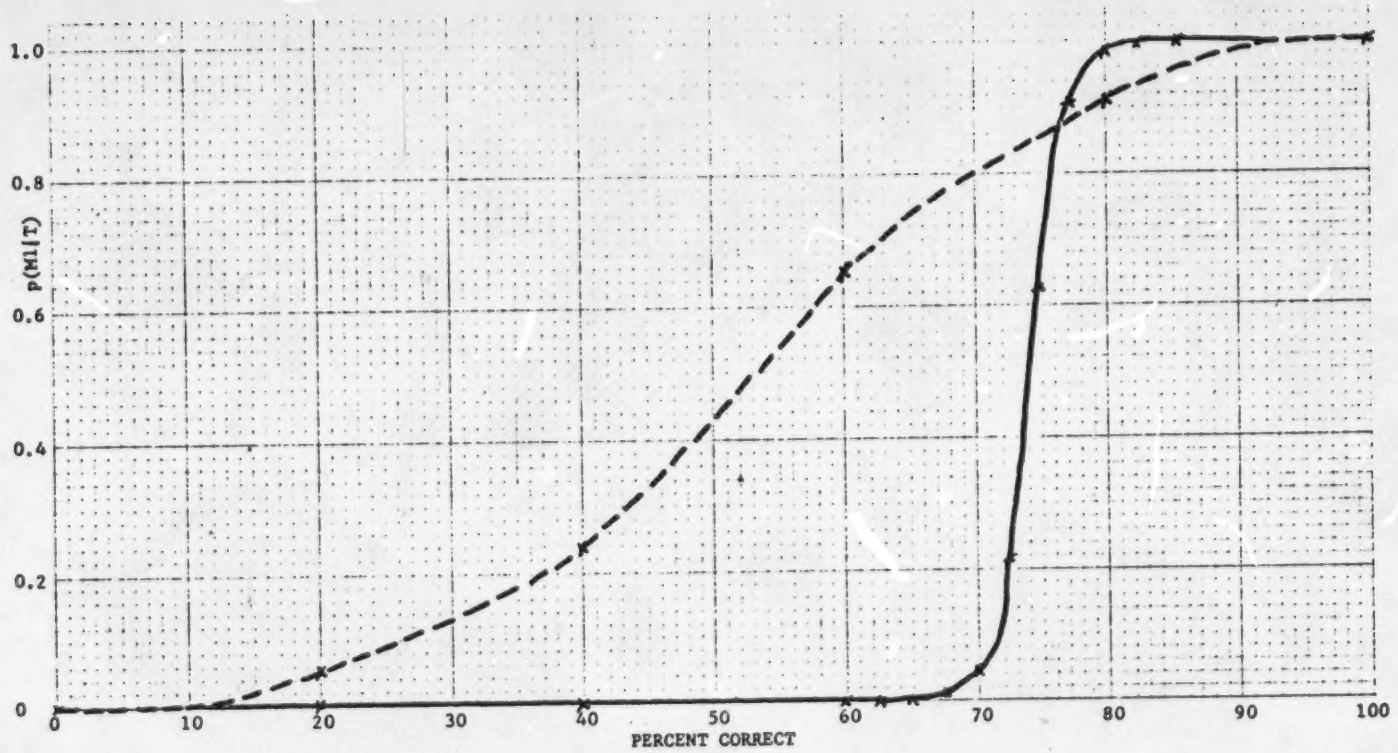
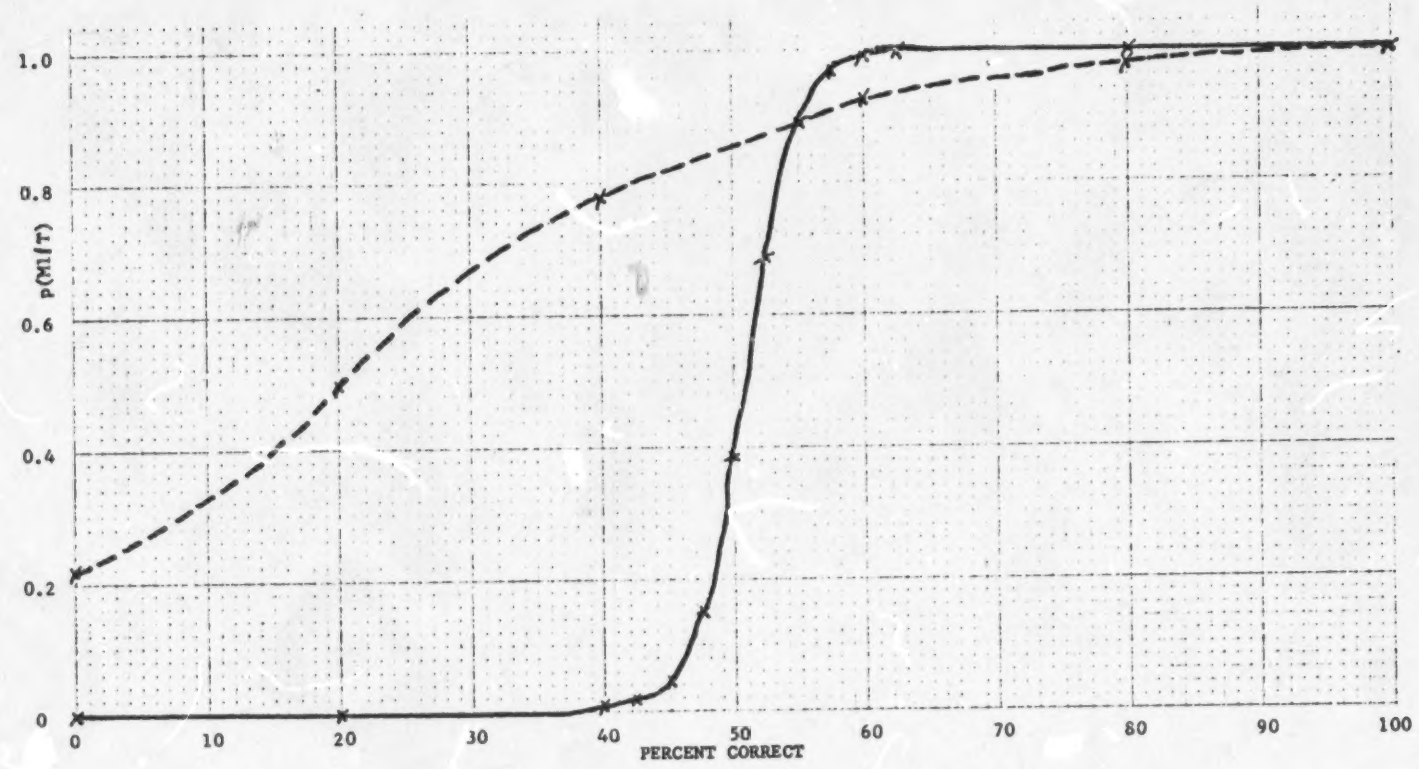


FIG. 9.

P(M1) = .9 P(M2) = .1

P(1|M1) = .7, P(1|M2) = .4

5 Item Test — — — — — 40 Item Test ————



626

24